Open-Sora: Democratizing Efficient Video Production for All

Year: 2024 | Citations: 441 | Authors: Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li

Abstract

Vision and language are the two foundational senses for humans, and they build up our cognitive ability and intelligence. While significant breakthroughs have been made in AI language ability, artificial visual intelligence, especially the ability to generate and simulate the world we see, is far lagging behind. To facilitate the development and accessibility of artificial visual intelligence, we created Open-Sora, an open-source video generation model designed to produce high-fidelity video content. Open-Sora supports a wide spectrum of visual generation tasks, including text-to-image generation, text-to-video generation, and image-to-video generation. The model leverages advanced deep learning architectures and training/inference techniques to enable flexible video synthesis, which could generate video content of up to 15 seconds, up to 720p resolution, and arbitrary aspect ratios. Specifically, we introduce Spatial-Temporal Diffusion Transformer (STDiT), an efficient diffusion framework for videos that decouples spatial and temporal attention. We also introduce a highly compressive 3D autoencoder to make representations compact and further accelerate training with an ad hoc training strategy. Through this initiative, we aim to foster innovation, creativity, and inclusivity within the community of AI content creation. By embracing the open-source principle, Open-Sora democratizes full access to all the training/inference/data preparation codes as well as model weights. All resources are publicly available at: https://github.com/hpcaitech/Open-Sora.